Using Performance Efficiency for Testing and Optimization of Visual Attention Models

Brian J. Stankiewicz, Nathan J. Anderson, Richard J. Moore

3M Company, 3M Center 235-2WN-64, St Paul, MN, USA 55144-1000

ABSTRACT

When developing a predictive tool for human performance one needs to have clear metrics to evaluate the model's performance. In the area of *Visual Attention Modeling* (VAM) one typically compares eye-tracking data collected on a group of human observers to the predictions made by a model. To evaluate the performance of these models one typically uses signal detection (Receiver Operating Characteristic (ROC)) that measures the predictive power of the system by comparing the model's predictions for an image to human eye tracking data. These ROC curves take into account the model's hit and false alarm rates and by averaging over a set of test images provides a final measure of the system's performance. In releasing a commercial visual attention system, we have spent considerable effort in developing metrics that allow for regression testing, that are useful for optimizing our visual attention model that takes into account the Upper-Theoretical Performance Limit for an image or classes of images. We describe how the Upper-Theoretical Performance Limit is calculated and how regression testing and parameter optimization benefit from this approach.

Keywords: image processing, visual attention modeling, regression testing

1. INTRODUCTION

Regression testing is critical to all software products. Often times there is a need to make sure that changes to the algorithms do not adversely affect the performance and accuracy of the software program. One of the challenges in doing regression testing and optimization predicting human behavior is that even under the best conditions there is inter-observer variability. In the area of Visual Attention Modeling researchers have measured human eye-fixations as a measure of what people are attending to in an image. To evaluate the performance of a model they then compare the model's predictions to that of the human behavior (i.e., the fixation data). In the case of eye fixations, there is a great deal of overlap between individuals, but there still is a certain amount of inter-observer variability and often times this variability will change on the types of images being evaluated (e.g., photographs vs. human generated graphics).

Since changes to the algorithm of a visual attention model can cause a pixel-by-pixel binary comparison of image output to fail, using a "golden standard" is inadequate for regression testing and model optimization. Instead, there is a is a need to develop metrics that are robust to slight changes in the predictions while indicating when changes to the algorithms causes significant degradation. Furthermore, regression tests need to validate all system output such as fixation regions, and viewing probabilities, heat maps, and stability to slight changes in the input. A robust metric would not only take into account how well the model is predicting human performance in

absolute terms, but it will also evaluate the model's performance relative to the variability for an image type and/or for a particular image. In this paper we describe the use of a *Performance* E*fficiency* measure that compares the model's performance to the **Upper-Theoretical Performance Limit** performance that takes into account the inter-observer variability. In the case of visual attention modeling, the Upper-Theoretical Performance Limit is bounded by the inter-observer (human) eye fixation variability for an image.

1.1 Visual Attention

The human visual system is capacity limited in a number of ways. To begin with, the human retina does not provide high resolution, high-color fidelity information across the entire retina. Instead, the center two-degrees of visual angle (called the *fovea*) are packed with small photoreceptors called cones that provide high resolution and color information. Outside of the fovea there are fewer cones and more *rod* photoreceptors that are used for low-light viewing conditions(Wandell, 1995). However, the human visual system has adapted to this heterogeneous photo mosaic by using rapid, saccadic eye fixations that rapidly moves the fovea from one region in the image to the next. These rapid saccades allow the human visual system to rapidly sample the scene with the high resolution fovea and over time, using visual memory, it "stitches" together a high resolution representation of the scene. It is well known that these fixations are not random (Tavassoli, 2009) and the initial fixations are well predicted by image properties such as color, motion, edges and contrast (Treisman, 1980). Because of the systematic selection of regions by the visual system based on image properties in the first few seconds of viewing an image, there has been a great deal of research in Vision Science and Computer Vision to develop theories and predictive computational models of human visual attention (Itti, 1998; Zhang, 2008). These models make explicit predictions about where people will initially look when viewing complex scenes such as shopping malls, streets, magazine pages, web pages, advertising content, etc. These models are typically based on the findings from behavior research in vision science (Itti, 1998) or based upon image statistics (Zhang, 2008).

2. PERFORMANCE MEASURES

Although these models are based upon fundamental findings in vision research there remains a certain amount of inter-subject variability and some unknown properties of how the visual system "decides" where to sample next. Given these sources of uncertainty, one would like to have an objective measure on how well a particular model, algorithm, or set of parameters is performing. One common way to evaluate visual attention models is to compare the model's performance to human eye-tracking data. Humans typically fixate their gaze (i.e., place their fovea) on the location that they are currently attending. Therefore, eye-fixation data serves as a surrogate to what people are actually attending to in an image. Eye-tracking equipment measures where an observer's center of gaze is as a function of time. However, models of visual attention do not make point predictions about where a fixation will occur, but instead provide a "heatmap" representation showing the most likely and least likely locations that the models predict (see left image in Figure 1). This makes comparing eye-fixations directly with a visual attention model heatmap a challenge.

One method that is typically used to evaluate the performance of a heatmap to the actual fixations is to use a Signal Detection *Response Operator Characteristics* (ROC) (Green, 1966) that take into account how well the model correctly predicts where human fixations will (hits) and will not occur (correct rejections) along with the incorrect predictions (false alarms and misses). ROC curves are

calculated by measuring the regions that the model correctly predicts a fixation will occur (hits) and incorrectly predicts (false alarms) across multiple thresholds of the model. Figure 1 and Figure 2 illustrate how one can take the output of a heatmap and generate progressively more liberal thresholds to generate more-and-more liberal predictions. The left image in Figure 1 illustrates a "heatmap" representation of the model's prediction where the "hotter" areas (reds) indicate that the more likely places that the model is predicting that a fixation will occur and the "cooler" colors (greens and blues) are less likely to receive visual attention. The right figure in Figure 1 shows an example of the thresholding maps where white indicates that the model's confidence is at or above a given threshold and the black regions are all of the areas that are below the model's threshold confidence. The illustration on the right side of Figure 2 shows an ROC curve that shows the predicted Hits X False Alarm rate as a function of the model's threshold value. By calculating the area under this curve one can generate an ROC value indicating the model's overall accuracy (Green, 1966). These values range between 0 and 1.0 with 1.0 being perfect performance and 0.5 being chance performance.

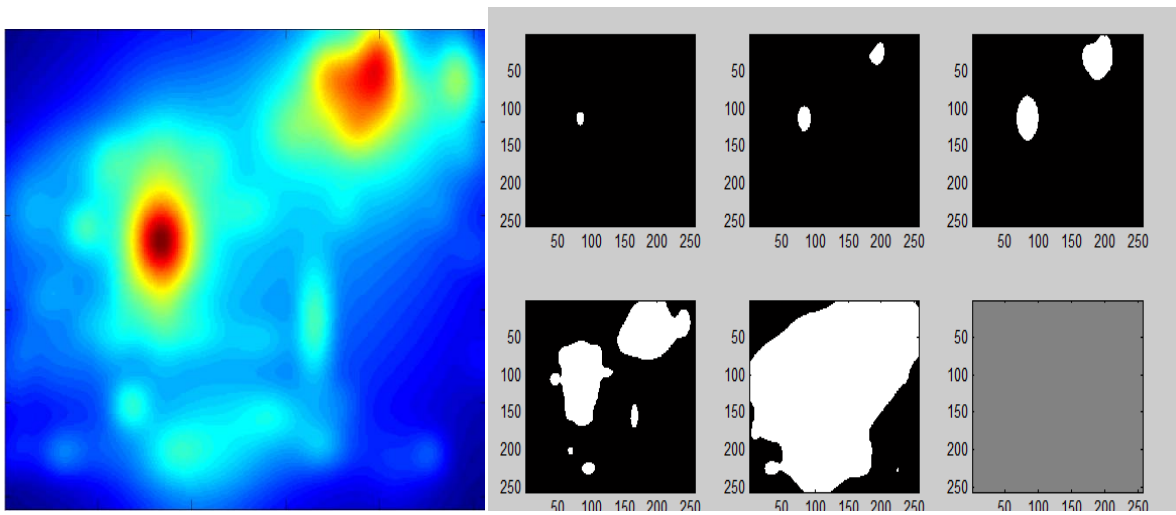

Figure 1. Illustration showing how the heatmap is thresholded to generate the different levels of prediction. The upper left image shows the "highest" values from the heatmap. Going from left-to-right and top-to-bottom we show increasingly lower threshold values

2.1 Upper-Theoretical Performance Limit

ROC calculations provide an objective, *absolute* performance measure for a particular model's image predictions. Although ROC values are objective and useful there is a particular weakness associated with them when evaluating human eye-fixation performance. That is they don't take into account the natural variation in the eye-tracking data (i.e., *between-subjects variability).*

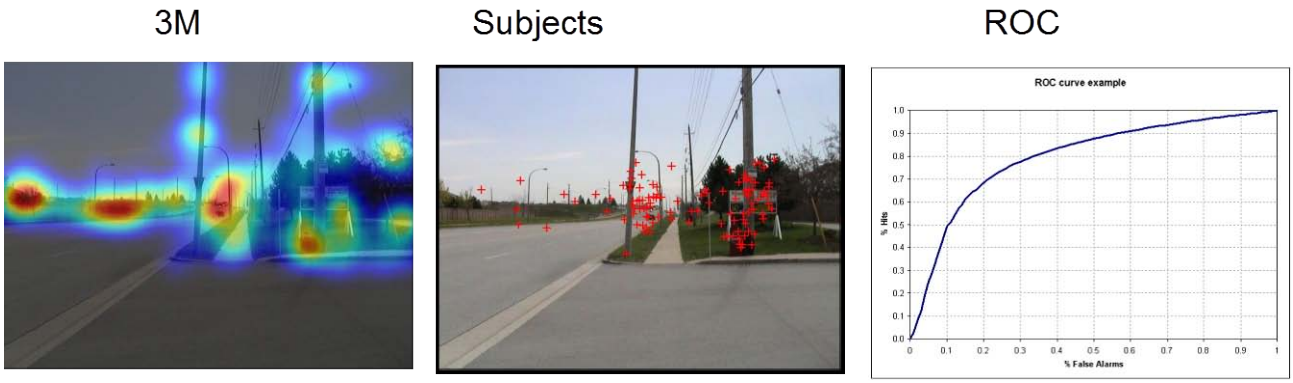

Figure 2. **Left:** Heatmap predictions generated by a visual attention model. **Center:** Eye-tracking data for multiple subjects collected on an image. **Right:** Example ROC curve for a Heat Map generated by the visual attention model. The Y-Axis measures the number of hits for a particular threshold while the X-Axis measures the number of False Alarms for that threshold value.

Typically when one runs an eye-tracking study they will collect data on a number of subjects and a number of images to try and reduce the amount of inter-observer variability. However, unlike most sampling theory problems predicting human eye-fixations has an interesting twist. In typical sampling theory problems, one generates more samples because there is "random" variation. In the case of human data and eye fixations, there is a certain amount of systematic variation that is generated based on the image type and even each individual image. That is, the variability between subjects will also vary as a function of image with some images having low inter-observer variability and others having high inter-observer variability. Without going into significant detail the inter-observer variability provides the **upper-theoretical performance limit** of any predictive model's performance. That is, the upper-theoretical performance limit for predicting eye-fixations is the ability of one visual system (one person or group of people) to predict the fixations of a second visual system (or group of people). Stated in a different way, one can not outperform a model that fully replicates the human visual system.

To measure the upper-theoretical performance limit we generated predictions using ½ of the eye-fixation data (where the subject groups were randomly selected) to generate a *Fixation Heat Map* One can think of this as using ½ of the subjects as an alternative "model" that measures the upper-theoretical performance limit. To generate the predictions we convolved a Gaussian kernel at each location where there was a fixation from ½ of the subjects. The Gaussian kernel was approximately 2-degree of visual angle (given the distance and size of the image used for the study), which corresponds roughly to the size of the fovea of the human eye (the high-resolution center 2-degrees of the retina) and within the resolution of the eye tracking systems used. The Right image in Figure 4 illustrates the output generated by this convolution and the left image in Figure 4 shows the actual fixation locations. We then used the generated Fixation Heatmap to predict the fixations from the second group of subjects using the ROC method described above. This ROC value provided us a measure of the Upper-Theoretical Performance Limit for each image.

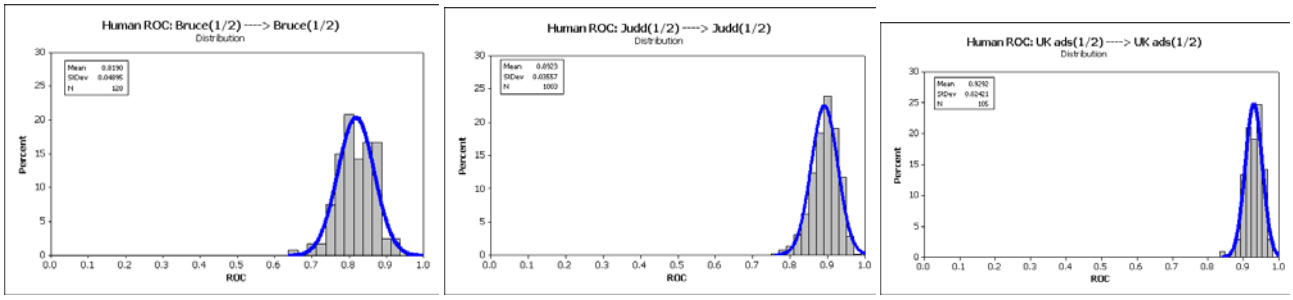

Figure 3. The distribution of Uppeer-Theoretical Performance values for the York University Data set (Mean ROC=0.82), MIT data set (Mean ROC=0.89) and the 3M data set (Mean ROC=0.93)

Figure 3 shows the distribution of Upper-Theoretical Performance Limit values for the different images for the for fixation data from the York University(Bruce, 2009) , MIT (Judd T., 2009) and data collected within 3M. We found that the Upper-Theoretical Performance Limit from the York University data set had an average ROC value of 0.82. The 3M replication study predicted the York data with an average ROC value of 0.81. This insignificant difference indicates that the methods and procedures used at 3M closely match those used at York University. The MIT data predicted itself with an average ROC value of 0.89 and the 3M advertising data predicted itself with an average ROC value of 0.93.

2.2 Prediction Efficiency

Because, theoretically any model of visual attention cannot outperform Upper-Theoretical Performance Limit (within the random noise range) we use this measure to provide an objective, *relative* performance measure. Specifically we were interested in measuring the model's **Prediction Efficiency** relative to the upper-theoretical performance limit.

Figure 4. **Left**: The combined fixations for 20 different participants who viewed this image in an eye-tracking study. **Right**: A *Fixation Heat Map* representing the variability of the participant's looked at this image.

In order to calculate the *Predictive Efficiency* (i.e., how well a visual attention model predicts eye-movements relative to the Upper-Theoretical Performance Limit), we calculated the ratio of the models performance to that of this theoretical limit (`Efficiency=100x[ROC(3MVAS)/ROC(Upper-Theor-Limit)]`). This measure provides an *efficiency* measure that indicates how well one can do

predicting eye-fixations using 3M VAS versus actually collecting eye-tracking data. When the efficiency value is close to 100% it means that the visual attention model is able to predict eye-fixations as well as an actual eye-tracking study.

2.3 Why use Prediction Efficiency?

One may wonder what the advantage is of using Prediction Efficiency rather than simply using ROC values alone. As a reminder, ROC values provide absolute measures for a model on a particular image, but it does not provide valuable information on how well the model is doing relative to the Upper-Theoretical Limit. For example, imagine collecting eye-fixation data on human subjects for an image of random intensity values. A visual attention model's ROC value would most likely approach chance predictive performance. With only the model's ROC values, one might incorrectly conclude that the model is performing poorly because it cannot accurately predict the human fixations. However, in this case there most likely won't be much consistency between human observers either. That is the Upper-Theoretical Performance Limit would also be close to chance. By calculating the Model's Prediction Efficiency by comparing the model's ROC to that of the Upper-Theoretical Limit, (~0.5/~0.5) one can see that the model is doing about as well as it can for those stimuli (i.e., the Prediction Efficiency would be approximately 100%).

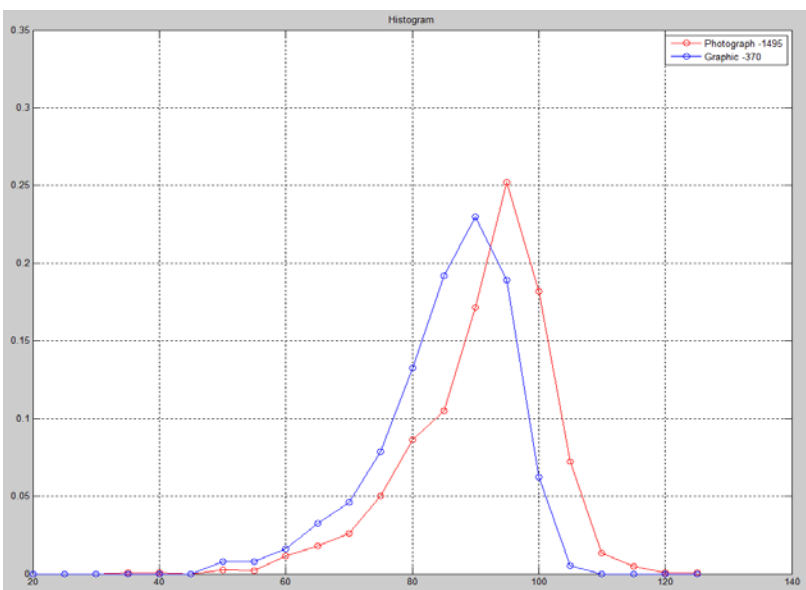

Figure 5. Visual attention efficiency distribution as a function of image type. The red line represents the histogram of efficiency values for photographs while the blue line indicates the efficiency values for graphics (e.g., advertisements and web pages).

When developing a computational model of visual attention one may see the need to understand which images and/or which image classes the model makes good versus better predictions. If one were to simply use the ROC approach, one might spend a great deal of effort trying to improve the model in places where there is little room for improvement (e.g., the random intensity images). By using the Prediction Efficiency one can begin to segregate the image set to identify the classes of images that have "room for improvement". As an example Figure 5 illustrates the Prediction Efficiency for a visual attention model for photographs versus graphics. As illustrated by this figure,

the model performs slightly better on photographs than on graphical (man-made) images. Interestingly enough, the model's ROC values are about the same for these two classes of images, however, the Upper-Theoretical Performance Limit is slightly different for these two classes of images.

3. SUMMARY & CONCLUSIONS

Visual attention models are complex with many inputs and parameters that can be manipulated and adjusted to give optimal performance. Using eye-tracking data we describe a method for evaluating *Predictive Efficiency* that measures how well the model's predictions are relative to the Upper-Theoretical Performance Limit—the human observer. This relative performance provides an objective measure for evaluating how much "head room" a particular image and/or a particular class of images has in terms of improving the model's performance.

Measuring the model's performance relative to the Upper-Theoretical Performance Limit allows one to focus their optimization efforts on images and/or classes where there is significant room for improvement. Furthermore, in regression testing, this approach allows for regression testing based on image type. Using Prediction Efficiency failures will occur only when the performance of the model is significantly below that of the Upper-Theoretical Performance Limit and not simply when the absolute performance is low.

4. REFERENCES

Bruce, N. D. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, , 9* (3), 1-24.

Green, D. S. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

Itti, C. K. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 20* (11), 1254-1259.

Judd T., E. K. (2009). Learning to Predict Where Humans Look. *IEEE International Conference on Computer Vision (ICCV)* .

Tavassoli, A. v. (2009). Eye movements selective for spatial frequency and orientation during active visual search. *Vision Research , 49*, 173-181.

Treisman, A. &. (1980). A feature-integration theory of attention. *Cognitive Psychology , 12*, 97-136.

Wandell, B. A. (1995). *Foundations of Vision.* Sunderland, Massachusetts: Sinauer Associates, Inc.

Zhang, L. T. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision , 8*, 1–20.